TIMESERIESANALYSIS

REPORT

GOLDFUTURESHISTORICALDATA

Introduction:

Gold futures data provides valuable insights into the dynamics of the gold market, a crucial component of the global economy and financial markets. As a commodity with intrinsic value, gold has historically served as a hedge against inflation, a safe-haven asset during times of economic uncertainty, and a store of value in diverse investment portfolios. Gold futures contracts, traded on commodity exchanges worldwide, offer investors and market participants a mechanism to speculate on the future price of gold, manage risk exposure, and engage in price discovery. Our project aims to explore and analyze historic gold futures data to uncover patterns, trends, and dependencies that shape the movements of gold prices over time. This study embarks on a comprehensive exploration of gold futures bond price time series forecasting, delving into the nuances of model fitting and predictive analytics. Through the application of cutting-edge methodologies such as ARIMA (AutoRegressive Integrated Moving Average), we aim to unlock the predictive potential inherent in gold futures bond price data.

By leveraging the power of time series analysis, we strive to unlock the hidden patterns and trends within the gold futures data, empowering decision-makers to navigate the complexities of the financial markets with confidence.

Problem Statement:

This project aims to leverage historical gold futures data to develop robust forecasting models that accurately predict future gold prices, enabling traders and investors to make informed decisions.

Data Source:

Dataset: Gold Futures Prices
Frequency: Daily
Period: [2020-04-24] to [2024-04-23]
Dataset link: https://in.investing.com/commodities/gold-historical-data

Data Description:

Gold futures data refers to the historical pricing and trading information related to gold futures contracts.

What are Gold Futures?

Gold Futures Contracts: These are standardized agreements to buy or sell a specified amount of gold at a predetermined price on a future delivery date. Futures contracts are traded on commodities exchanges.

Purpose: Gold futures allow investors, speculators, and producers (such as mining companies) to hedge against the risk of price fluctuations in the gold market.

Contract Specifications: Each gold futures contract specifies the quantity of gold, quality (purity), delivery date, and delivery location.

Gold Futures Data Includes:

1. Date: The trading date of the gold futures contract.
2. Price: The closing price of the gold futures contract for the day.
3. Open: The opening price of the gold futures contract for the day.
4. High: The highest price reached by the gold futures contract during the day.
5. Low: The lowest price reached by the gold futures contract during the day.
6. Volume: The number of contracts traded during the day.
7. Change Percent: The percentage change in the closing price from the previous trading day.

We will be using the “Price” column, which is the closing price of the Gold Futures Bond,against the “Date” column, as our time series data for analysis.

Exploratory Data Analysis:

- ***TIME SERIES PLOT :***

To interpret the above time series plot we will perform decomposition of the time series and plot the decomposition components.

- ***DECOMPOSITION PLOT :***

The **first plot** shows **original** gold future prices over time. The plot displays the actual daily prices of gold futures. It shows how the variable of interest changes over time. There's a clear upward trend starting around the early 2023, which indicates increasing values in the series.

The **second plot** represents the **trend component** after decomposition. This component captures the long-term progression of the series. It appears that there is a relatively stable pattern until early 2023, after which there's a noticeable **upward trend**. This could represent a change in the underlying fundamentals driving the series value upwards.

The **third plot** displays the **seasonal component** after decomposition.
The seasonal plot shows the repeating short-term cycle in the data. The regular up-and-down spikes indicate consistent seasonal fluctuations within each year. The amplitude (height) of these seasonal spikes appears to be quite stable over time, which suggests that the seasonal effect **does not change** much from year to year.
The **fourth plot** displays the **residual component**. The residual plot shows what is left after removing the trend and seasonal components. Ideally, the residual plot should not exhibit any clear patterns or trends. In this case, the residual plot appears to have some **random fluctuations**, which is a good indication that the trend and seasonal components have captured

most of the data's information.Any remaining patterns or trends in the residual plot might indicate that the model did not fully capture all aspects of the data.

- ***TRENDANALYSIS***
  - ***LINEAR TREND MODEL***

The equation for the linear trend model is **Yt = 1755.57+0.2302*t** and the associated performance measures are:
Mean Absolute Percentage Error (MAPE): 4.2
Mean Absolute Deviation (MAD): 79.3
Mean Squared Deviation (MSD): 11118.5

Mean Absolute Percentage Error (MAPE):MAPE measures the average percentage difference between the actual and predicted values. A lower MAPE indicates better accuracy.In this case MAPE suggests that, on average, the model's predictions are within 4.2% of the actual values.

Mean Absolute Deviation (MAD):
MAD measures the average absolute difference between the actual and predicted values. It provides a measure of the model's accuracy without considering the direction of errors. In this case MAD of 79.3 indicates that, on average, the model's predictions deviate from the actual values by approximately 79.3 units.

Mean Squared Deviation (MSD):
MSD measures the average squared difference between the actual and predicted values. It gives more weight to larger errors compared to MAD.A higher MSD value of 11118.5 suggests that the model has larger errors compared to MAD.

Overall, the model seems to have reasonably good accuracy based on the MAPE and MAD values. However, the higher MSD suggests that there may be outliers or cases where the model's predictions deviate significantly from the actual values.

- ***QUADRATIC TREND MODEL***

The equation for the quadratic trend model is **Yt= 1894.04-0.5656*t+0.000763*t^2**
and the associated performance measures:
Mean Absolute Percentage Error (MAPE): 3.55
Mean Absolute Deviation (MAD): 66.18
Mean Squared Deviation (MSD): 7305.80

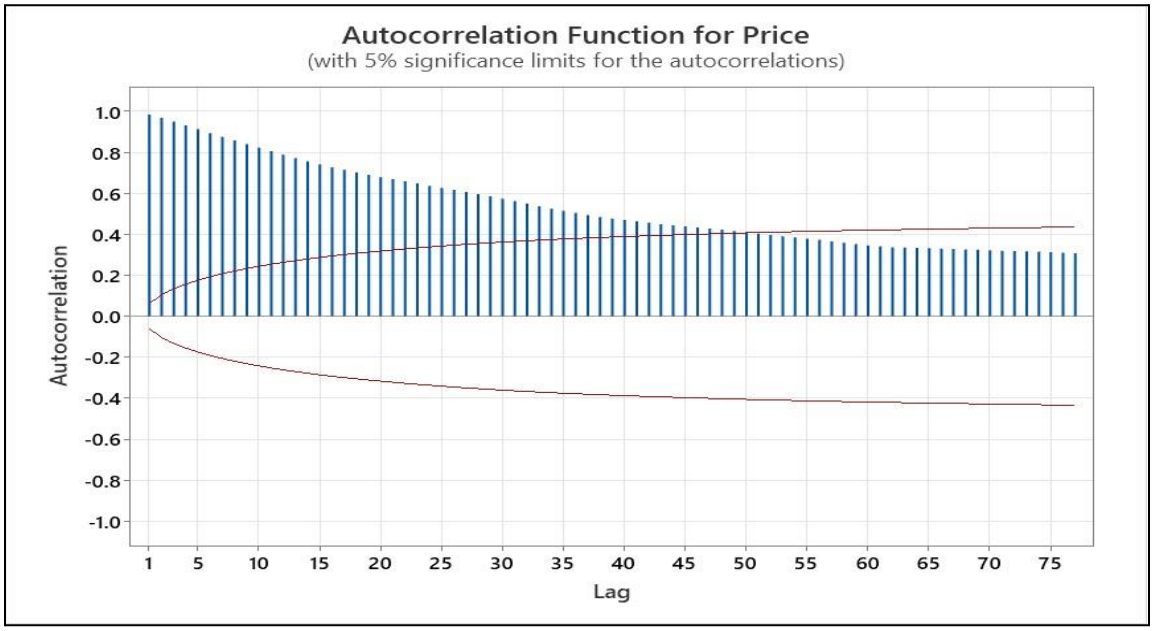
Mean Absolute Percentage Error (MAPE):
In this case, a MAPE of 3.55% suggests that, on average, the model's predictions are within 3.55% of the actual values.
Mean Absolute Deviation (MAD):
A MAD of 66.18 indicates that, on average, the model's predictions deviate from the actual values by approximately 66.18 units.

Mean Squared Deviation (MSD):
A lower MSD value of 7305.80 compared to the previous linear trend suggests that the model has smaller errors overall and is more accurate in terms of squared deviations.

The low MAPE and MAD values indicate that the quadratic trend model has good accuracy and is able to predict the time series values with relatively small errors.

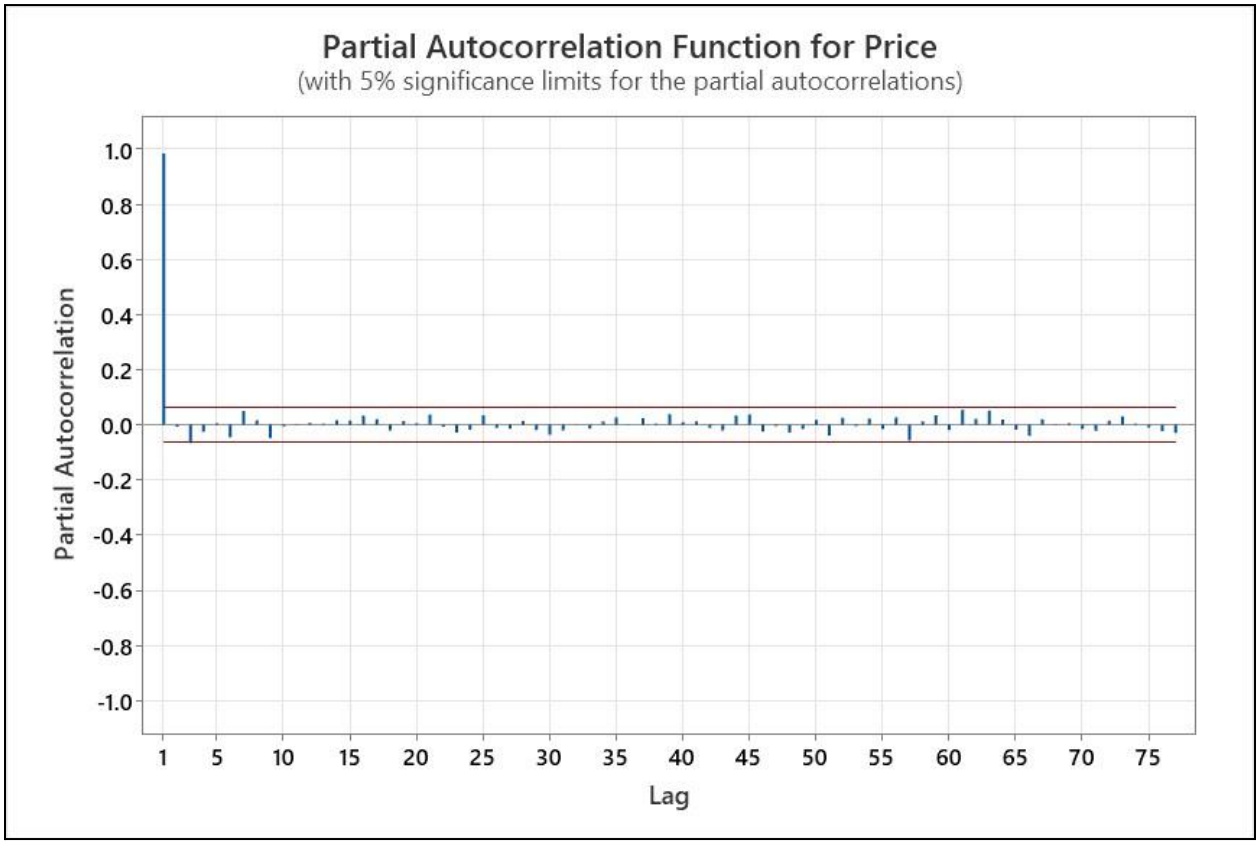
The lower MSD compared to the previous linear trend model suggests that the quadratic trend model performs better in terms of minimizing squared errors and capturing the variability in the data. Thus we can say that the data has a **Quadratic Upwards Trend.**

- ***AUTOCORRELATION FUNCTION (ACF) PLOT :***



The ACF plot shows the autocorrelation coefficients for different lag values (lags) of the time series.Autocorrelation measures the correlation between a time series and its lagged versions. The above ACF declines gradually from 1 to 0 over a prolonged period of time, and there are around 50 significant lags identified from the plot, thus the data is **not stationary**.

- ***PARTIAL AUTOCORRELATION FUNCTION (PACF) PLOT :***



The partial autocorrelation at **lag 1** is notably above the significance line, which suggests that there's a statistically significant correlation between each observation and the one immediately preceding it. The bar at lag 1 extends beyond the blue significance limits, indicating that the correlation is not due to chance. The fact that all other lags are within the significance limits suggests that there are no other autoregressive terms that are significantly correlated with the time series once the effect of the first lag has been accounted for.

- ***DICKEY-FULLER TEST ON ORIGINAL DATA :***

  The Dickey-Fuller test helps assess whether a time series is stationary or non-stationary.

Augmented Dickey-Fuller Test for Price

Method

Maximum lag order for terms in the regression model	21
Criterion for selecting lag order	Minimum AIC
Additional terms	Constant
Selected lag order	0
Rows used	1042

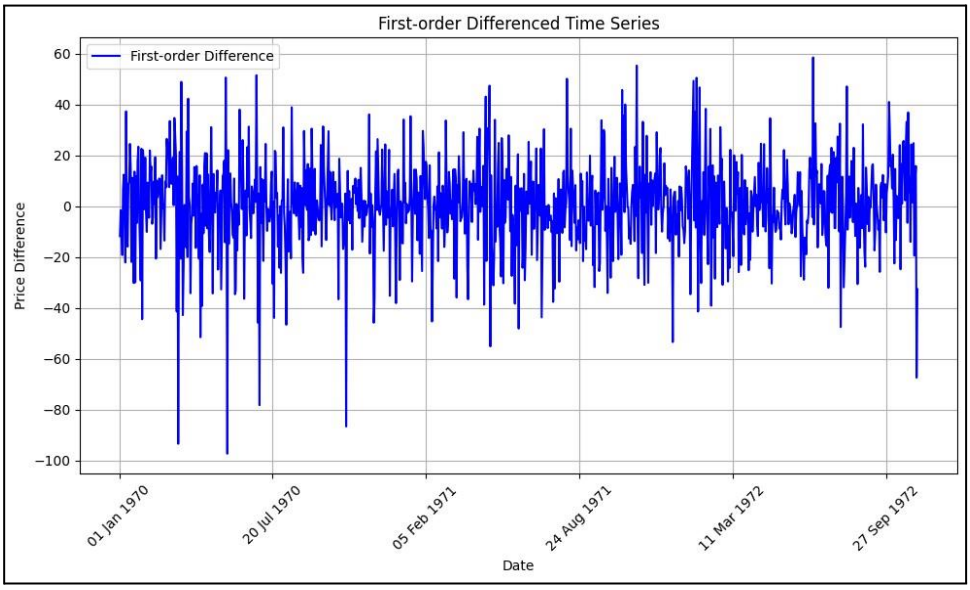
Augmented Dickey-Fuller Test

Null hypothesis:	Data are non-stationary
Alternative hypothesis:	Data are stationary

Test Statistic	P-Value	Recommendation
-1.08009	0.723	Test statistic > critical value of -2.86432. Significance level = 0.05 Fail to reject null hypothesis. Consider differencing to make data stationary.

Applying the Dickey-Fuller test it tells us that our original data is not stationary and that differencing is necessary to make data stationary.

- ***TIME SERIES PLOT OF THE FIRST ORDER DIFFERENCED DATA :***



T

The first order differencing is done to make the data stationary.

From the graph we can interpret that the mean is constant around 0 and the variance is also mostly constant. So from the plot, we can say that the first order differenced series is stationary. Next we apply the Augmented Dickey-Fuller test to validate that the series is stationary.

Now applying the Augmented Dickey-Fuller test again on the differenced data we have,

Augmented Dickey-Fuller Test for Differenced Data

Method

Maximum lag order for terms in the regression model	21
Criterion for selecting lag order	Minimum AIC
Additional terms	Constant
Selected lag order	0
Rows used	1041
Rows unused	1

Augmented Dickey-Fuller Test

Null hypothesis:	Data are non-stationary
Alternative hypothesis:	Data are stationary

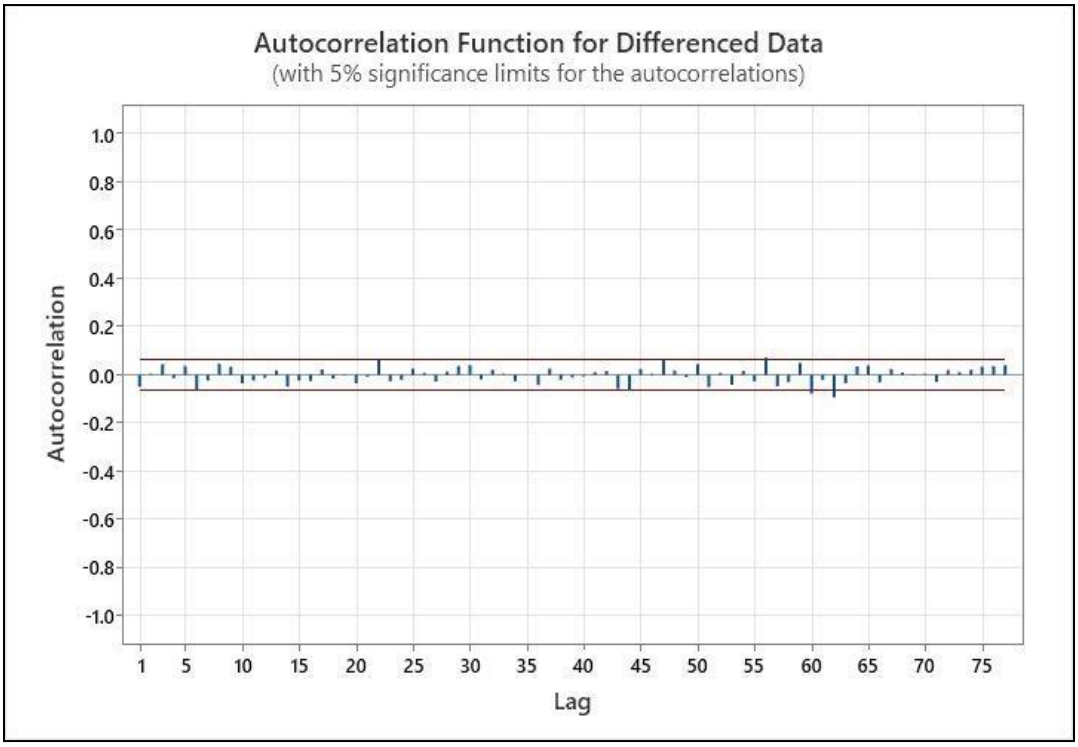
Test Statistic	P-Value	Recommendation
-32.6420	0.000	Test statistic <= critical value of -2.86432. Significance level = 0.05 Reject null hypothesis. Data appears to be stationary, not supporting differencing.

Thus, the Augmented Dickey-Fuller test verifies that the **differenced series is stationary**, and thus we will continue with our analysis using the stationary differenced series only.

- ***ACF PLOT OF THE DIFFERENCED DATA :***

Autocorrelations

Lag	ACF	T	LBQ
1	-0.0493674	-1.59	2.54
2	0.0042382	0.14	2.56
3	0.0443596	1.43	4.62
4	-0.0150514	-0.48	4.86
5	0.0372248	1.20	6.31
6	-0.0624540	-2.00	10.40
7	-0.0237886	-0.76	11.00
8	0.0459077	1.47	13.21
9	0.0335660	1.07	14.40
10	-0.0354577	-1.13	15.72

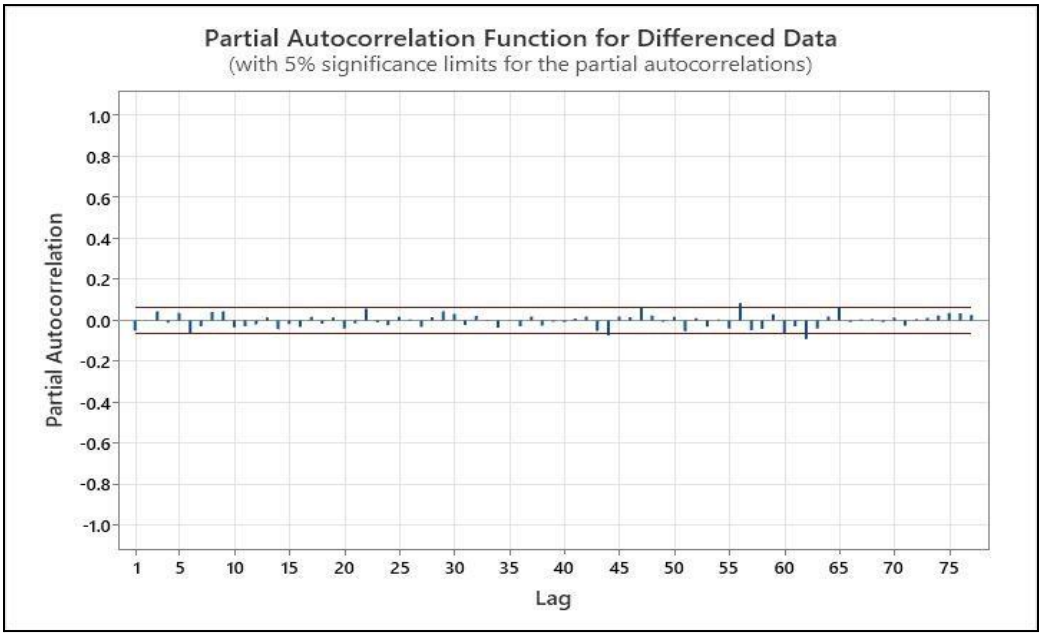


The ACF values show weak to moderate autocorrelation in the time series, with no strong long-term dependencies but occasional significant autocorrelations at specific lags. Here we can interpret that the values are dependent on observations at lag1 i.e on the previous date.

- ***PACF PLOT OF THE DIFFERENCED DATA :***

Partial Autocorrelations

Lag	PACF	T
1	-0.0493674	-1.59
2	0.0018055	0.06
3	0.0447668	1.44
4	-0.0107164	-0.35
5	0.0357555	1.15
6	-0.0611204	-1.97
7	-0.0290911	-0.94
8	0.0408951	1.32
9	0.0448523	1.45
10	-0.0328699	-1.06



The PACF plot value at lag 1 indicates a direct relationship between the current date and the price at lag 1.

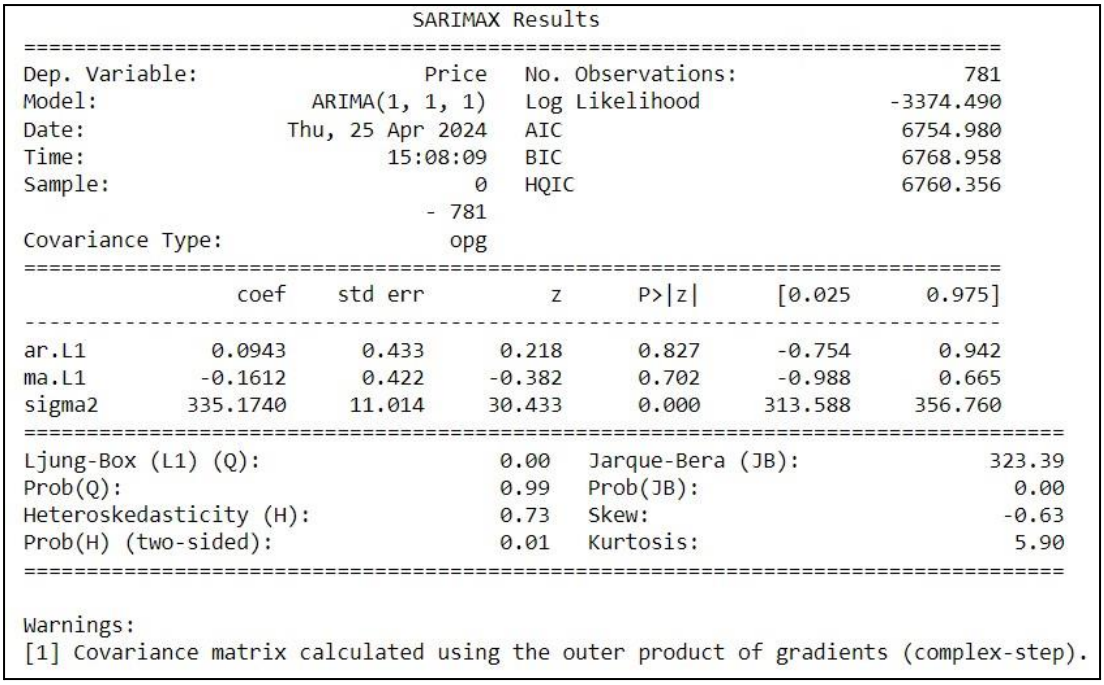
MODELING THE TIME SERIES DATA :

We fit different models to our data to check which model fits best and thus can give us our best forecast values.

ARIMA MODELS :

We had divided the data into training and test data of 75% and 25% each. We then built the model on the train data and tested the model accuracy and diagnostics using the test data. Following are the ARIMA Models we compared:

- ***ARIMA (1,1,1)***



SARIMAX Results

Dep. Variable:	Price	No. Observations:	781
Model:	ARIMA(1, 1, 1)	Log Likelihood	-3374.490
Date:	Thu, 25 Apr 2024	AIC	6754.980
Time:	15:08:09	BIC	6768.958
Sample:	0	HQIC	6760.356
	- 781		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	0.0943	0.433	0.218	0.827	-0.754	0.942
ma.L1	-0.1612	0.422	-0.382	0.702	-0.988	0.665
sigma2	335.1740	11.014	30.433	0.000	313.588	356.760

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	323.39
Prob(Q):	0.99	Prob(JB):	0.00
Heteroskedasticity (H):	0.73	Skew:	-0.63
Prob(H) (two-sided):	0.01	Kurtosis:	5.90

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Root Mean Squared Error (RMSE) on test data: 118.47154738868495

Coefficients Interpretation:

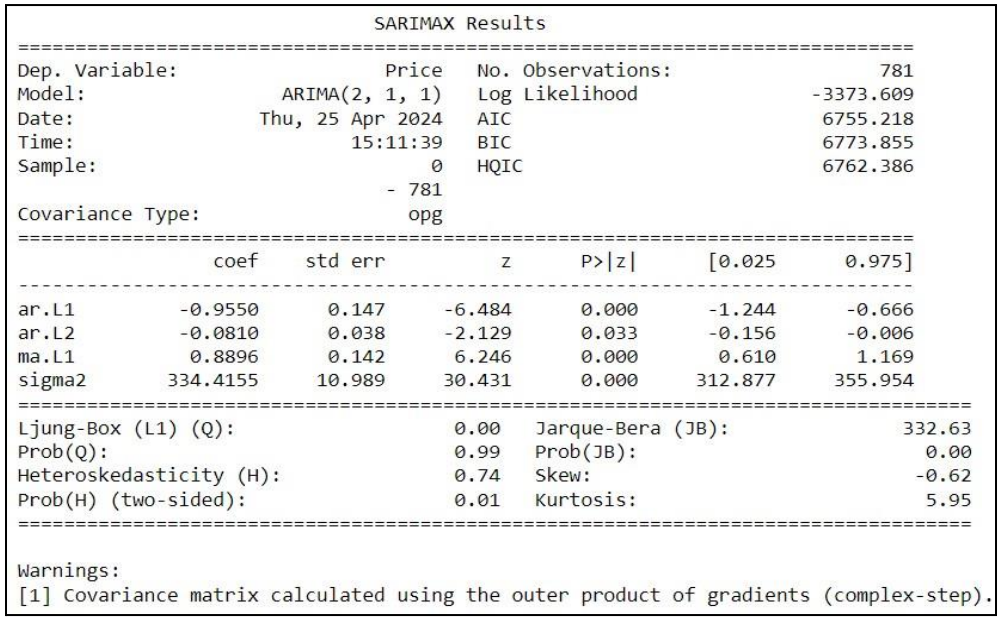
- The AR(1) coefficient (ar.L1) is 0.0943 with a p-value of 0.827, indicating it's not statistically significant. This suggests weak linear dependence between current and lagged prices.
- The MA(1) coefficient (ma.L1) is -0.1612 with a p-value of 0.702, also not statistically significant. This implies the error at the previous time step doesn't strongly influence the current price.

Diagnostic Tests Interpretation:

- The Ljung-Box test shows a p-value of 0.99, indicating no evidence of autocorrelation at lag 1 in the residuals.
- The Jarque-Bera test has a p-value of 0.00, suggesting that the residuals may not be normally distributed, possibly indicating a limitation of the model.

AIC value : 6754.980 **BIC value :** 6768.958

- ***ARIMA (2,1,1)***



```
                               SARIMAX Results
==============================================================================
Dep. Variable:                  Price   No. Observations:                  781
Model:                 ARIMA(2, 1, 1)   Log Likelihood               -3373.609
Date:                Thu, 25 Apr 2024   AIC                           6755.218
Time:                        15:11:39   BIC                           6773.855
Sample:                             0   HQIC                          6762.386
                                  - 781
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.9550      0.147     -6.484      0.000      -1.244      -0.666
ar.L2         -0.0810      0.038     -2.129      0.033      -0.156      -0.006
ma.L1          0.8896      0.142      6.246      0.000       0.610       1.169
sigma2       334.4155     10.989     30.431      0.000     312.877     355.954
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):               332.63
Prob(Q):                              0.99   Prob(JB):                         0.00
Heteroskedasticity (H):               0.74   Skew:                            -0.62
Prob(H) (two-sided):                  0.01   Kurtosis:                         5.95
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

Root Mean Squared Error (RMSE) on test data: 118.40617480505693

Coefficients Interpretation:

- The AR(1) coefficient (ar.L1) is -0.9550 with a p-value of 0.000, indicating statistical significance. This suggests a strong negative linear relationship between the current price and its lagged value.
- The AR(2) coefficient (ar.L2) is -0.0810 with a p-value of 0.033, also statistically significant. This implies a weaker negative linear relationship with the price two periods ago.
- The MA(1) coefficient (ma.L1) is 0.8896 with a p-value of 0.000, indicating statistical significance. This suggests a strong positive linear relationship between the error at the previous time step and the current price.

Diagnostic Tests Interpretation:

- The Ljung-Box test shows a p-value of 0.99, indicating no evidence of autocorrelation at lag 1 in the residuals.
- The Jarque-Bera test has a p-value of 0.00, suggesting that the residuals may not be normally distributed, potentially indicating a limitation of the model.

AIC value : 6755.218 **BIC value :** 6773.855

- ***ARIMA (1,1,2)***

```
                               SARIMAX Results
==============================================================================
Dep. Variable:                  Price   No. Observations:                  781
Model:                 ARIMA(1, 1, 2)   Log Likelihood               -3373.679
Date:                Thu, 25 Apr 2024   AIC                           6755.357
Time:                        15:12:03   BIC                           6773.994
Sample:                             0   HQIC                          6762.525
                                  - 781
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.8775      0.152     -5.786      0.000      -1.175      -0.580
ma.L1          0.8139      0.152      5.346      0.000       0.516       1.112
ma.L2         -0.0791      0.037     -2.147      0.032      -0.151      -0.007
sigma2       334.4731     10.967     30.497      0.000     312.978     355.969
===================================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):               333.84
Prob(Q):                              0.97   Prob(JB):                         0.00
Heteroskedasticity (H):               0.74   Skew:                            -0.62
Prob(H) (two-sided):                  0.01   Kurtosis:                         5.95
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
```

Root Mean Squared Error (RMSE) on test data: 118.42456672301799

Coefficients Interpretation:

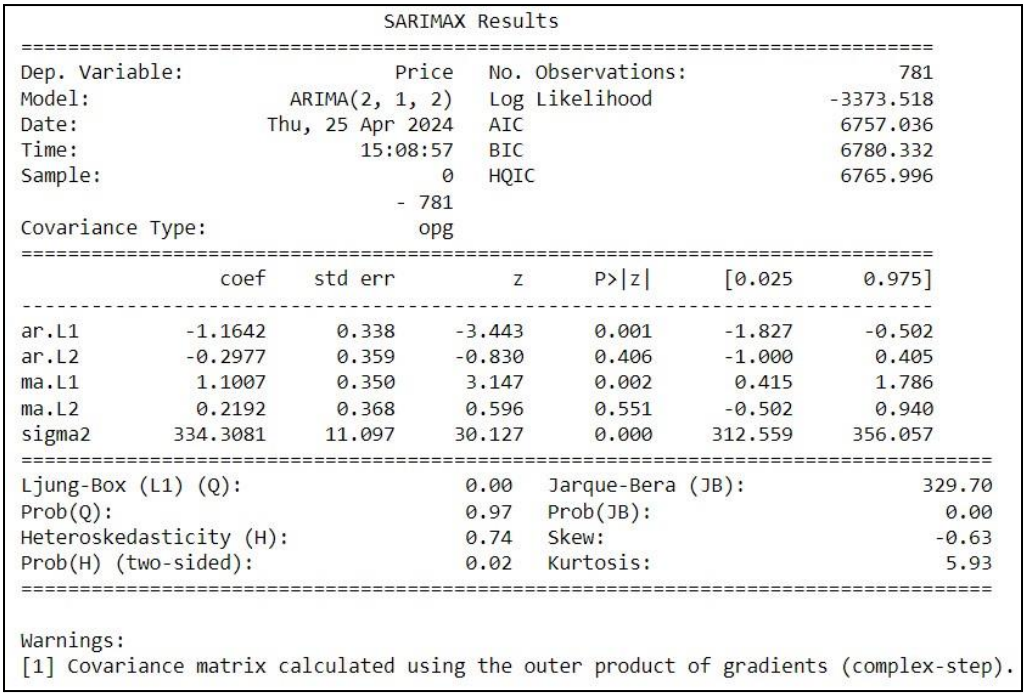
- The AR(1) coefficient (ar.L1) is -0.8775 with a p-value of 0.000, indicating statistical significance. This suggests a strong negative linear relationship between the current price and its lagged value.
- The MA(1) coefficient (ma.L1) is 0.8139 with a p-value of 0.000, indicating statistical significance. This suggests a strong positive linear relationship between the error at the previous time step and the current price.
- The MA(2) coefficient (ma.L2) is -0.0791 with a p-value of 0.032, indicating statistical significance. This implies a weak negative linear relationship with the error two periods ago.

Diagnostic Tests Interpretation:

- The Ljung-Box test shows a p-value of 0.97, indicating no evidence of autocorrelation at lag 1 in the residuals.
- The Jarque-Bera test has a p-value of 0.00, suggesting that the residuals may not be normally distributed, potentially indicating a limitation of the model.

AIC value : 6755.357 **BIC value :** 6773.994

- ***ARIMA (2,1,2)***



SARIMAX Results

Dep. Variable:	Price	No. Observations:	781
Model:	ARIMA(2, 1, 2)	Log Likelihood	-3373.518
Date:	Thu, 25 Apr 2024	AIC	6757.036
Time:	15:08:57	BIC	6780.332
Sample:	0	HQIC	6765.996
	- 781		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-1.1642	0.338	-3.443	0.001	-1.827	-0.502
ar.L2	-0.2977	0.359	-0.830	0.406	-1.000	0.405
ma.L1	1.1007	0.350	3.147	0.002	0.415	1.786
ma.L2	0.2192	0.368	0.596	0.551	-0.502	0.940
sigma2	334.3081	11.097	30.127	0.000	312.559	356.057

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	329.70
Prob(Q):	0.97	Prob(JB):	0.00
Heteroskedasticity (H):	0.74	Skew:	-0.63
Prob(H) (two-sided):	0.02	Kurtosis:	5.93

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Root Mean Squared Error (RMSE) on test data: 118.36993274853327

Coefficients Interpretation:

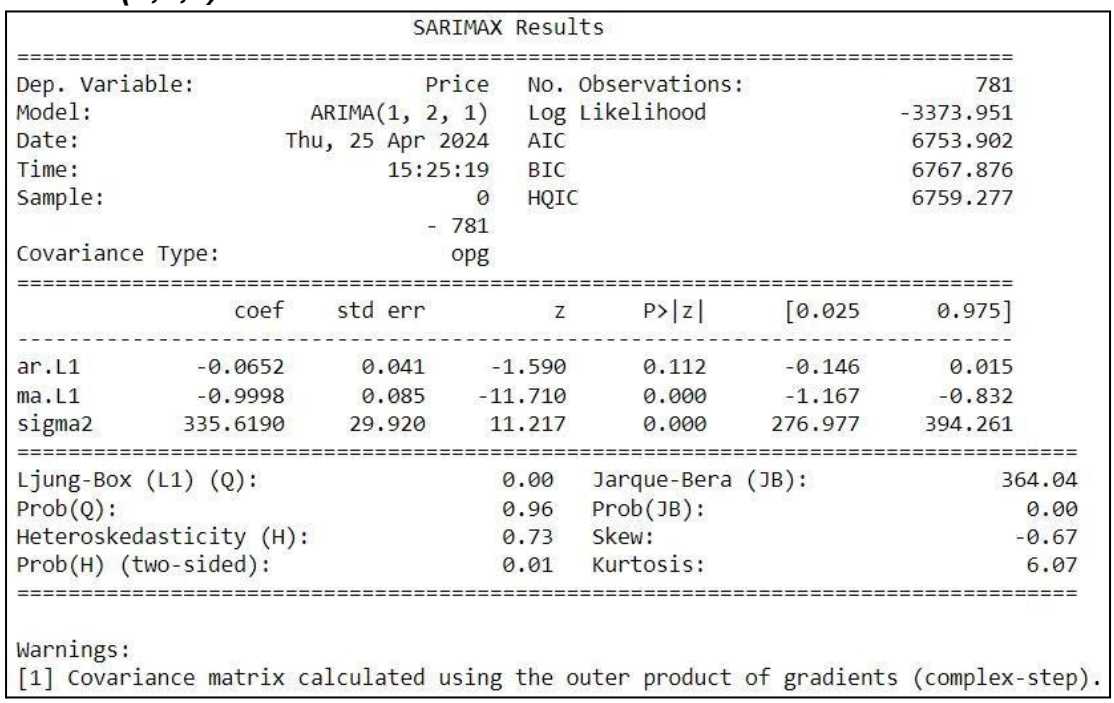
- The AR(1) coefficient (ar.L1) is -1.1642 with a p-value of 0.001, indicating statistical significance. This suggests a strong negative linear relationship between the current price and its lagged value.
- The AR(2) coefficient (ar.L2) is -0.2977 with a p-value of 0.406, indicating it is not statistically significant. This implies the second lagged value might not contribute significantly to the model.
- The MA(1) coefficient (ma.L1) is 1.1007 with a p-value of 0.002, indicating statistical significance. This suggests a strong positive linear relationship between the error at the previous time step and the current price.
- The MA(2) coefficient (ma.L2) is 0.2192 with a p-value of 0.551, indicating it is not statistically significant. This suggests the second lagged error might not contribute significantly to the model.

Diagnostic Tests Interpretation:

- The Ljung-Box test shows a p-value of 0.97, indicating no evidence of autocorrelation at lag 1 in the residuals.
- The Jarque-Bera test has a p-value of 0.00, suggesting that the residuals may not be normally distributed, potentially indicating a limitation of the model.

AIC value : 6757.036 **BIC value :** 6780.332

- ***ARIMA (1,2,1)***



SARIMAX Results

Dep. Variable:	Price	No. Observations:	781
Model:	ARIMA(1, 2, 1)	Log Likelihood	-3373.951
Date:	Thu, 25 Apr 2024	AIC	6753.902
Time:	15:25:19	BIC	6767.876
Sample:	0	HQIC	6759.277
	- 781		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.0652	0.041	-1.590	0.112	-0.146	0.015
ma.L1	-0.9998	0.085	-11.710	0.000	-1.167	-0.832
sigma2	335.6190	29.920	11.217	0.000	276.977	394.261

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	364.04
Prob(Q):	0.96	Prob(JB):	0.00
Heteroskedasticity (H):	0.73	Skew:	-0.67
Prob(H) (two-sided):	0.01	Kurtosis:	6.07

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Root Mean Squared Error (RMSE) on test data: 97.78725934774141

Coefficients Interpretation:

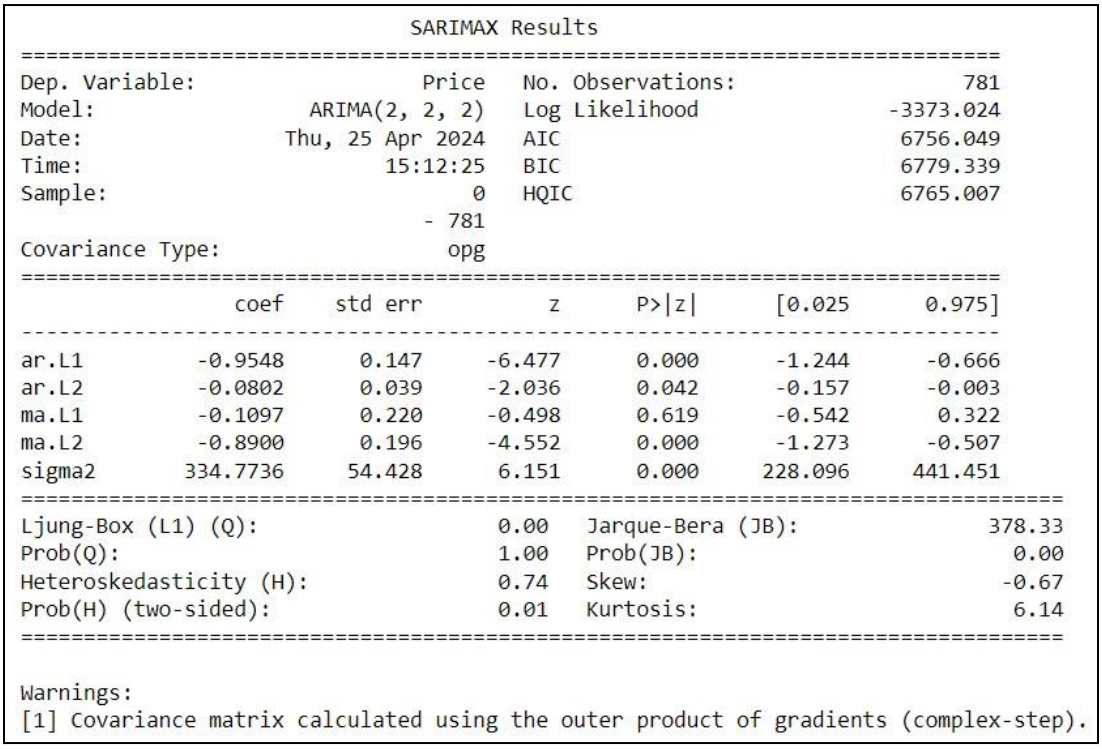
- The AR(1) coefficient (ar.L1) is -0.0652 with a p-value of 0.112, indicating no statistical significance.
- The MA(1) coefficient (ma.L1) is -0.9998 with a p-value of 0.000, indicating statistical significance. This suggests a strong negative linear relationship between the error at the previous time step and the current price.

Diagnostic Tests Interpretation:

- The Ljung-Box test shows a p-value of 0.96, indicating no evidence of autocorrelation at lag 1 in the residuals.
- The Jarque-Bera test has a p-value of 0.00, suggesting that the residuals may not be normally distributed, potentially indicating a limitation of the model.

AIC value : 6755.357 **BIC value** : 6773.994

- ***ARIMA (2,2,2)***



SARIMAX Results

Dep. Variable:	Price	No. Observations:	781
Model:	ARIMA(2, 2, 2)	Log Likelihood	-3373.024
Date:	Thu, 25 Apr 2024	AIC	6756.049
Time:	15:12:25	BIC	6779.339
Sample:	0	HQIC	6765.007
	- 781		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.9548	0.147	-6.477	0.000	-1.244	-0.666
ar.L2	-0.0802	0.039	-2.036	0.042	-0.157	-0.003
ma.L1	-0.1097	0.220	-0.498	0.619	-0.542	0.322
ma.L2	-0.8900	0.196	-4.552	0.000	-1.273	-0.507
sigma2	334.7736	54.428	6.151	0.000	228.096	441.451

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	378.33
Prob(Q):	1.00	Prob(JB):	0.00
Heteroskedasticity (H):	0.74	Skew:	-0.67
Prob(H) (two-sided):	0.01	Kurtosis:	6.14

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

Root Mean Squared Error (RMSE) on test data: 97.81975715220406

Coefficients Interpretation:

- The AR(1) coefficient (ar.L1) is -0.9548 with a p-value of 0.000, indicating statistical significance. This suggests a strong negative linear relationship between the current price and its lagged value.
- The AR(2) coefficient (ar.L2) is -0.0802 with a p-value of 0.042, indicating statistical significance. This implies a weak negative linear relationship with the price two periods ago.
- The MA(1) coefficient (ma.L1) is -0.1097 with a p-value of 0.619, indicating it is not statistically significant. This suggests that the first lagged error may not contribute significantly to the model.
- The MA(2) coefficient (ma.L2) is -0.8900 with a p-value of 0.000, indicating statistical significance. This suggests a strong negative linear relationship with the error two periods ago.

Diagnostic Tests Interpretation:

- The Ljung-Box test shows a p-value of 1.00, indicating no evidence of autocorrelation at lag 1 in the residuals.
- The Jarque-Bera test has a p-value of 0.00, suggesting that the residuals may not be normally distributed, potentially indicating a limitation of the model.

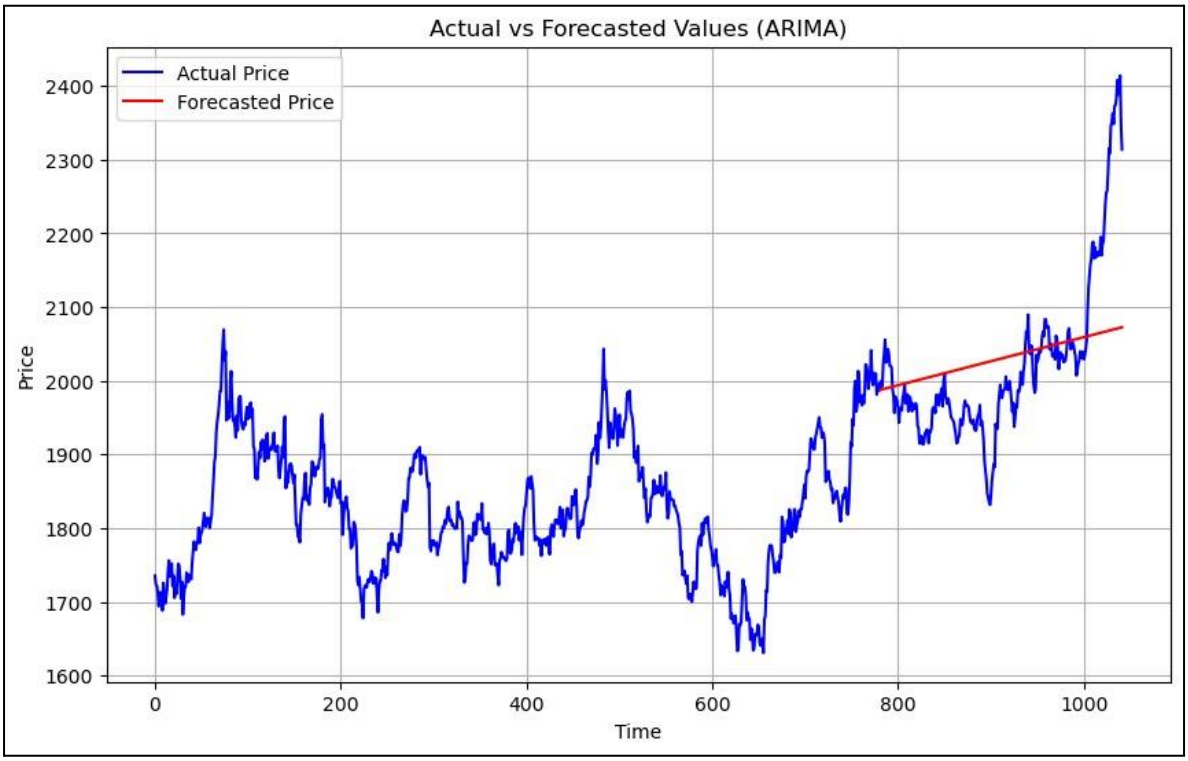
AIC value : 6756.049 **BIC value :** 6779.339

Comparing the models :

ARIMA	P-Value					AIC	BIC	RMSE
	Constant	AR		MA				
		1	2	1	2			
(1,1,1)	0.000	0.827		0.702		6754.98	6768.958	118.471
(2,1,1)	0.000	0.000	0.033	0.000		6755.218	6773.855	118.406
(1,1,2)	0.000	0.000		0.000	0.032	6755.357	6773.994	118.424
(2,1,2)	0.000	0.001	0.406	0.002	0.551	6757.036	6780.332	118.37
(1,2,1)	0.000	0.112		0.000		6753.902	6767.876	97.787
(2,2,2)	0.000	0.000	0.042	0.619	0.000	6756.049	6779.339	97.82

On comparing the 6 ARIMA Models, we can see that ARIMA(1,2,1) has 2 significant coefficients and the other has a lower intensity of insignificance, and the model has the lowest AIC and BIC and also the lowest RMSE among all other models. Thus, we can say that for our dataset, ARIMA(1,2,1) model fits best.

This is how the actual values of the test data vs forecast values plot looks :



HOLT WINTERS EXPONENTIAL SMOOTHING :

We also checked how well Holt's Winter Exponential Smoothing fits our data. For this also we divided the data into 75% training and 25% test sets.

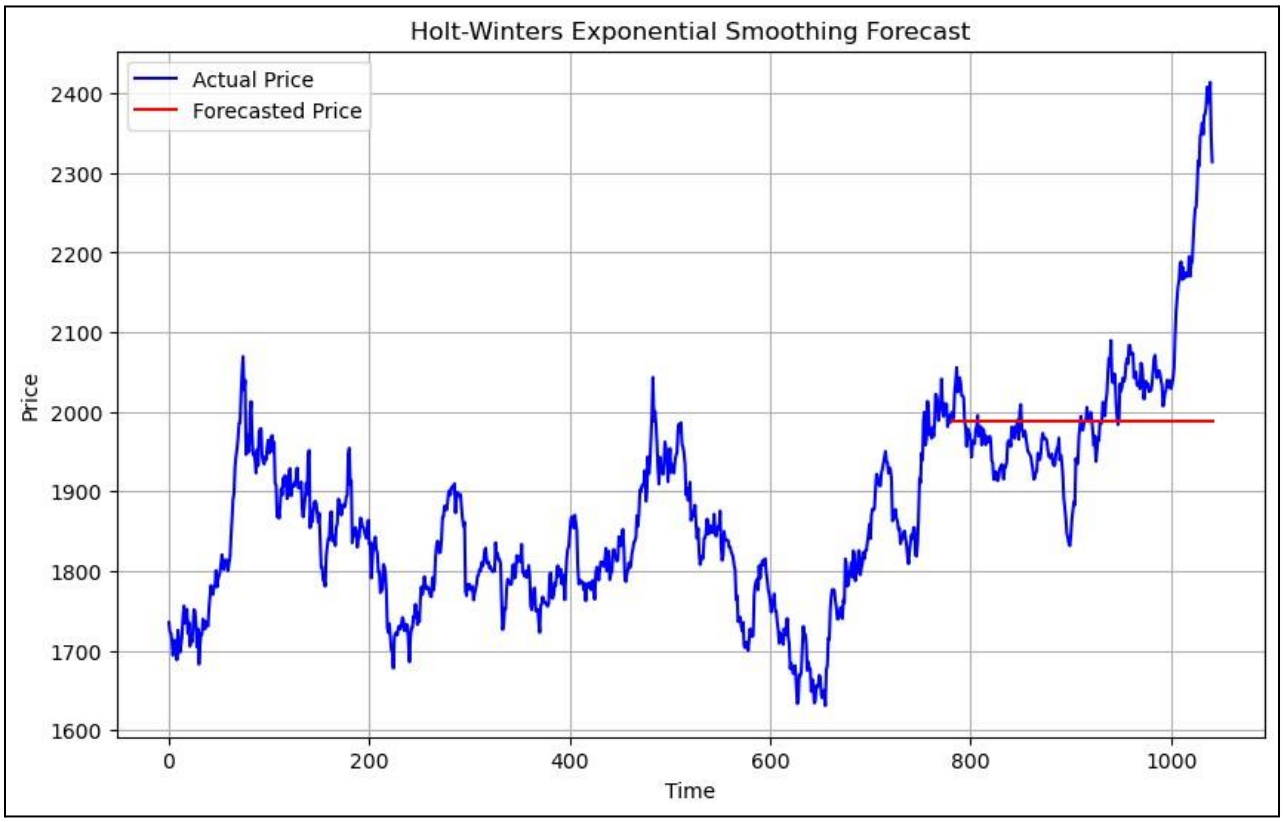
ExponentialSmoothing Model Results

Dep. Variable:	Price	No. Observations:	781
Model:	ExponentialSmoothing	SSE	261449.540
Optimized:	True	AIC	4544.282
Trend:	None	BIC	4553.603
Seasonal:	None	AICC	4544.334
Seasonal Periods:	None	Date:	Thu, 25 Apr 2024
Box-Cox:	False	Time:	19:36:13
Box-Cox Coeff.:	None		

	coeff	code	optimized
smoothing_level	0.9325595	alpha	True
initial_level	1734.7943	l.0	True

Root Mean Squared Error (RMSE) on test data: 118.47526241148918

Following is the plot of actual vs forecast values of the model :



Though the AIC and BIC of this model is much lesser than our suggested ARIMA(1,2,1) model, the RMSE is significantly larger than that of the ARIMA model and the above plot looks like the forecast values are constant after a time-point, which shows that it is not that good a model for our data.

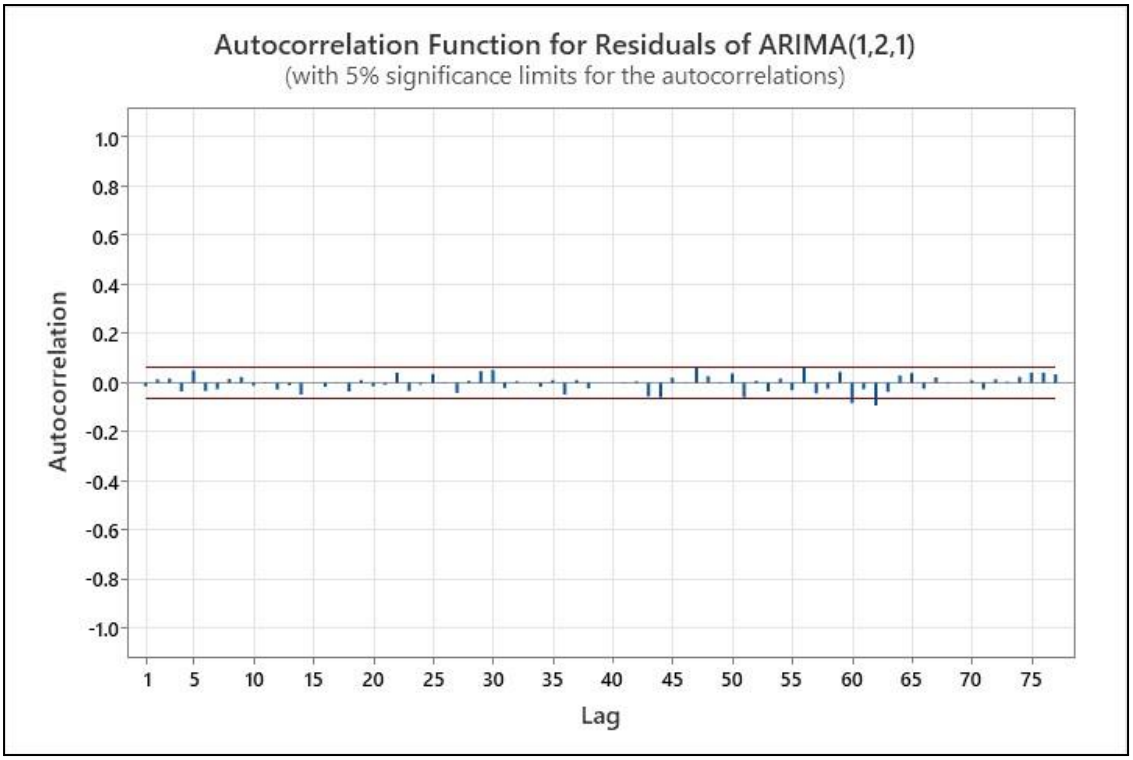
So we choose ARIMA (1,2,1) to be a better fit compared to the Holt Winters Smoothed model.

MODEL DIAGNOSTICS

- ***Ljung-Box Test :***

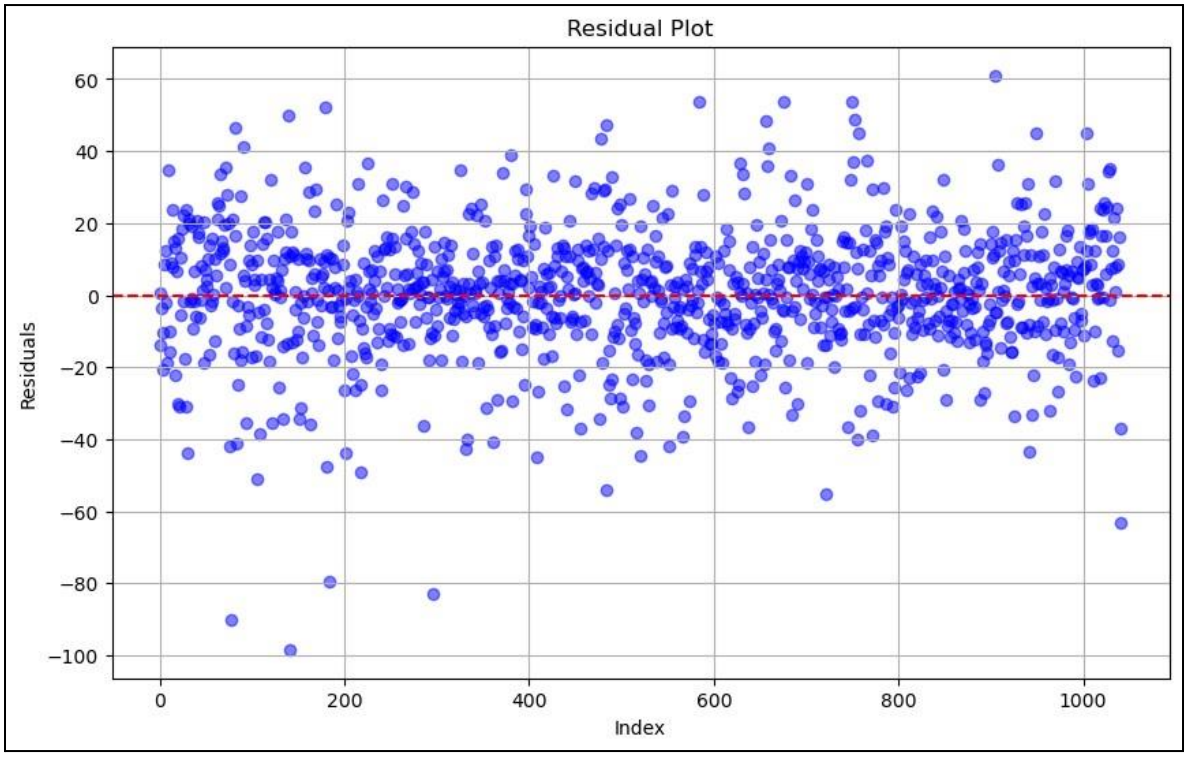
  The Ljung-Box test shows a p-value of 0.97, thus we fail to reject its null hypothesis which indicates no evidence of autocorrelation at lag 1 in the residuals, which is our desired result.

- ***Autocorrelation Function (ACF) Plot of the Residuals :***



  The ACF graph shows that there is no autocorrelation between the residuals at any significant lags.

- ***Residual Plot (for checking Homoscedasticity in the residuals) :***

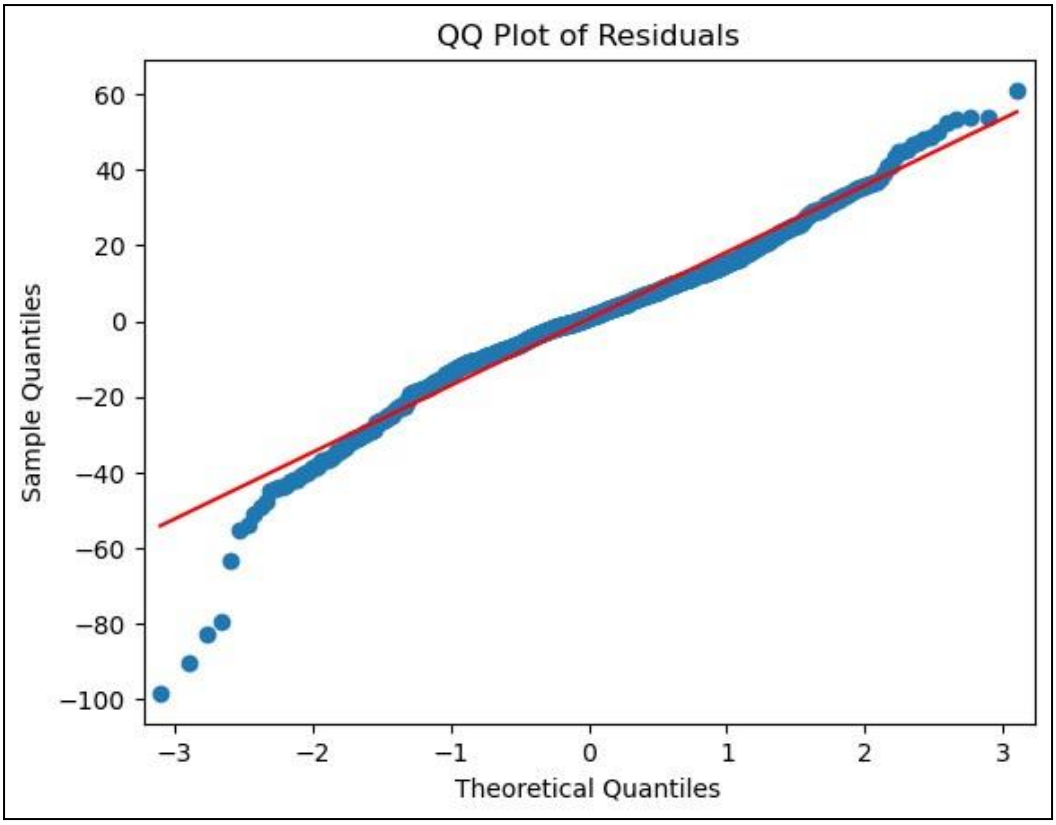


If the plot is such that the residuals can be contained in a horizontal band fashion (and residual fluctuates more or less in a random manner inside the band), then there are no visible model defects. This clearly indicates that the errors are homoscedastic in nature.

- ***Jarque-Bera Test :***

  The Jarque-Bera test shows a p-value of 0.00, thus we reject its null hypothesis, which indicates that the residuals do not follow normal distribution.

- ***QQ Plot or Normal Probability Plot:***



This plot has upward and downward curves at both extremes. This indicates that the underlying distribution is heavy-tailed, i.e., the tails of the underlying distribution are thicker than the tails of normal distribution. Thus we can say that the errors in the model are not normally distributed.

Thus, after performing all diagnostic checks, we can say that the residuals in our model are stationary, i.e, they are white noise. But since they don't follow Normal Distribution, we can't say they are Gaussian White Noise.

Tools used for the project : Minitab and Python

FINAL CONCLUSION :

Summarizing our findings, we would like to state that there is a quadratic upward trend in our data, and no evidence of any seasonal or cyclical variations in it. As for the best ARIMA model that fits it, we would say that ARIMA (1,2,1) is the best fit for the data. However, there can be other better advanced time series

models which fits the data more accurately than ARIMA, however, in the ARIMA family, ARIMA(1,2,1) is possibly the best model for this data.

---------- X ----------